



Name - Utkrist Raj Garg

College – Lakshmi Narain College of Technology

Group – 9

Project – 2

Email - [utkristraj2820@gmail.com](mailto:utkristraj2820@gmail.com)

TOPIC

Heart Disease Analysis and Prediction

Contents

[ABSTRACT 6](#_Toc71759208)

[IMPORTS 7](#_Toc71759209)

[Importing libraries 7](#_Toc71759210)

[Importing our Dataset 7](#_Toc71759211)

[CHECKING OUR DATA 8](#_Toc71759212)

[Checking the type of Data our DataFrame consists 8](#_Toc71759213)

[Checking our ‘df’ DataFrame 8](#_Toc71759214)

[COLUMN INFORMATION 9](#_Toc71759215)

[Numerical columns 9](#_Toc71759216)

[Categorical columns 9](#_Toc71759217)

[Working on the Numerical Columns 10](#_Toc71759218)

[Sorting 10](#_Toc71759219)

[Mean 10](#_Toc71759220)

[Variance 11](#_Toc71759221)

[Standard Deviation 11](#_Toc71759222)

[Median 12](#_Toc71759223)

[Skewness 13](#_Toc71759224)

[Correlation of numerical columns with target column 14](#_Toc71759225)

[Heatmap 15](#_Toc71759226)

[EDA of Numerical Columns 16](#_Toc71759227)

[Code block 16](#_Toc71759228)

[Age column 17](#_Toc71759229)

[Resting Blood Pressure 18](#_Toc71759230)

[Cholesterol 19](#_Toc71759231)

[Maximum heart rate achieved 20](#_Toc71759232)

[ST depression rate 21](#_Toc71759233)

[EDA of Categorical Columns 22](#_Toc71759234)

[Code Block 22](#_Toc71759235)

[Gender 23](#_Toc71759236)

[Age with target 24](#_Toc71759237)

[Fasting blood sugar 25](#_Toc71759238)

[Exercise induced angina 26](#_Toc71759239)

[Chest Pain 27](#_Toc71759240)

[Resting electrocardiographic result 28](#_Toc71759241)

[ST Segment 29](#_Toc71759242)

[Major blood vessels 30](#_Toc71759243)

[Thalassemia 31](#_Toc71759244)

[Relation b/w Numerical & Categorical Column w.r.t Target 32](#_Toc71759245)

[Code block 32](#_Toc71759246)

[Age and Numerical columns 33](#_Toc71759247)

[Fasting blood sugar with numerical columns 35](#_Toc71759248)

[Exercise induced angina with numerical columns 37](#_Toc71759249)

[Chest pain with numerical columns 39](#_Toc71759250)

[Resting electrographic with numerical columns 41](#_Toc71759251)

[Thalassemia with numerical columns 43](#_Toc71759252)

[Data Preprocessing 45](#_Toc71759253)

[Machine learning prediction on the Dataset 46](#_Toc71759254)

[Decision Tree 46](#_Toc71759255)

[K-Nearest Neighbors 47](#_Toc71759256)

[Comparing Accuracy between our KNN and Decision Tree Classifier 49](#_Toc71759257)

[Conclusion 50](#_Toc71759258)

# ABSTRACT

I am creating a Data Analysis Project on Heart Disease Prediction. The project uses raw data in form of a .csv file and transforms into Data Analysis. This project is an attempt of data analyzing Heart Disease Prediction with the help of data science and data analytics in python code. Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of clinical data analysis. The amount of data in the healthcare industry is huge. Data mining turns the large collection of raw healthcare data into information that can help to make informed decisions and predictions.

Heart Disease (CHD) is the most common type of disease, killing over 370,000 people annually. This makes heart disease a major concern to be dealt with. But it is difficult to identify heart disease because of several risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate, and many other factors. Because of these factors, scientists have turned towards modern approaches like Data Mining and Machine Learning for predicting the disease.

This analysis will help us find the basis behind common notions of about the causes of heart disease from purely a dataset perspective.

# IMPORTS

## Importing libraries



Numpy is used creating arrays.

Pandas for creating and manipulating Data Frames and for depth analysis.

Matplotlib or creating canvas and plot our visualization.

Seaborn for creating different graphs to visually analyze our dataset.

%matplotlib inline allows us to see visualize our graphs in the notebook

## Importing our Dataset



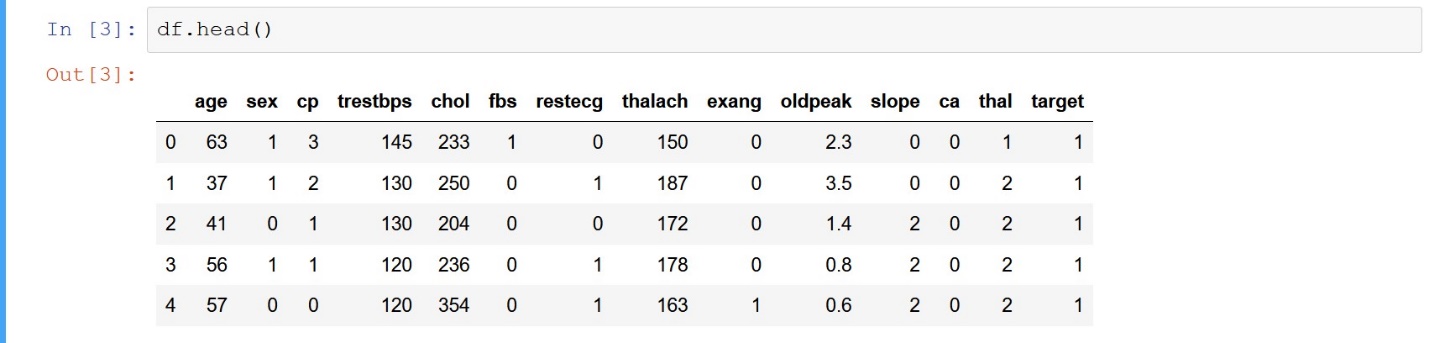
Initializing our dataset as ‘df’ using pandas inbuilt read\_csv function.

# CHECKING OUR DATA

## Checking the type of Data our DataFrame consists

The info() function tells us about the type of data in our columns in the dataset.

## Checking our ‘df’ DataFrame



The head() function displays 5 rows form the top by default but can display n rows. We verify the data types in the DataFrame.

# COLUMN INFORMATION

## Numerical columns

age – the person’s age in years.

trestbps – the person’s resting blood pressure.

chol –  The person's cholesterol measurement in mg/dl.

thalach – The person’s maximum heart rate achieved.

oldpeak – ST depression induced by exercise relative to rest.

## Categorical columns

sex – person’s sex where 1 = male & 2 = female.

fbs – fasting blood pressure >120mg/dl where 1 = True & 2 = False.

exang – exercise induced angina where 1 = True & 2 = False.

cp – chest pain where 0: typical angina, 1: atypical angina, 2: non-anginal pain, 3: asymptomatic

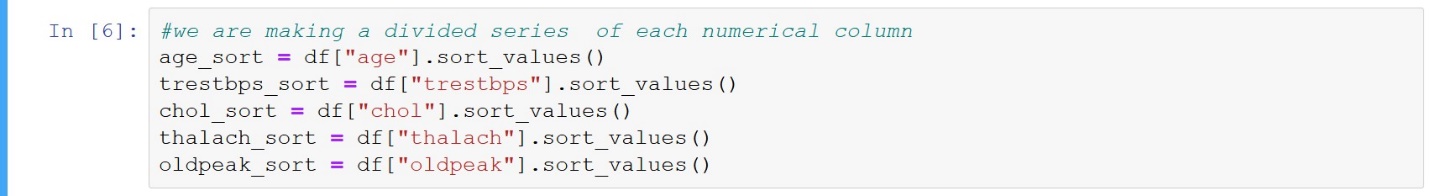
restecg – Resting electrocardiographic measurement 0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria.

slope –  the slope of the peak exercise ST segment where 0: upsloping, 1: flat, 2: downsloping)

thal –  A blood disorder called thalassemia where 0= normal; 1 = fixed defect; 2 = eversible defect: 3 = Irreversable defect.

# Working on the Numerical Columns

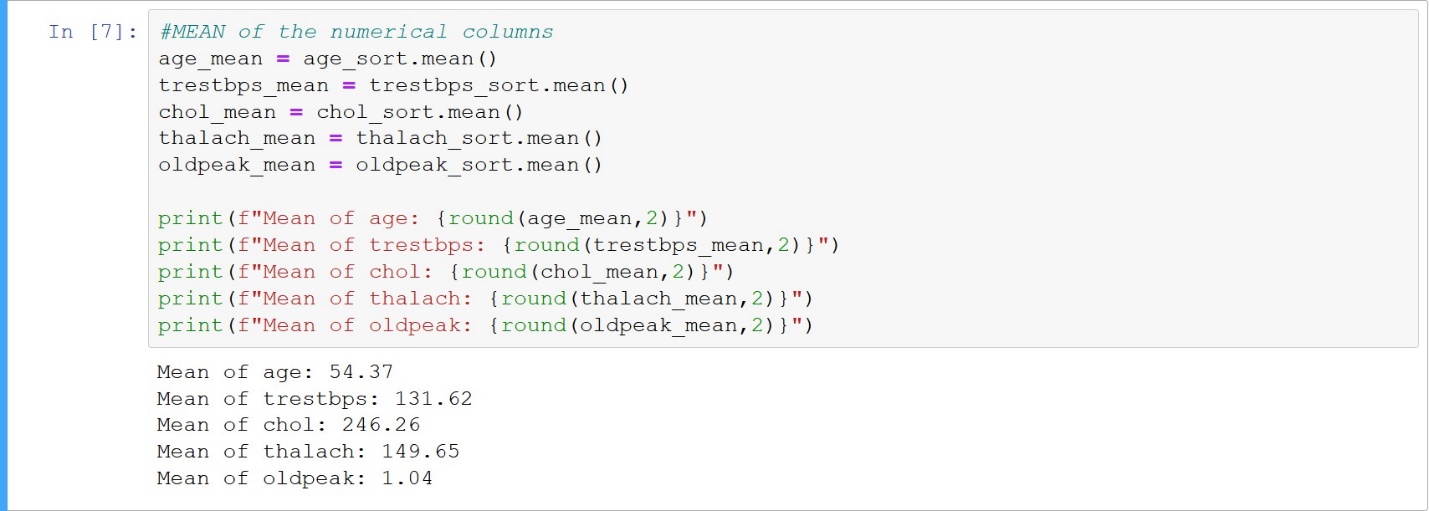
## Sorting



The sort\_values() function will sort the data in ascending order. Here, we sort the data per column.

After sorting the data per columns we assign that data in a variable like, age data is sorted and assigned to the age\_sort variable making it an individual series. We do this same for every numerical column.

## Mean



Now we that sorted series to find the mean of every column.

The .mean() function gives the mean of every individual sorted series.

Then we print that mean accordingly.

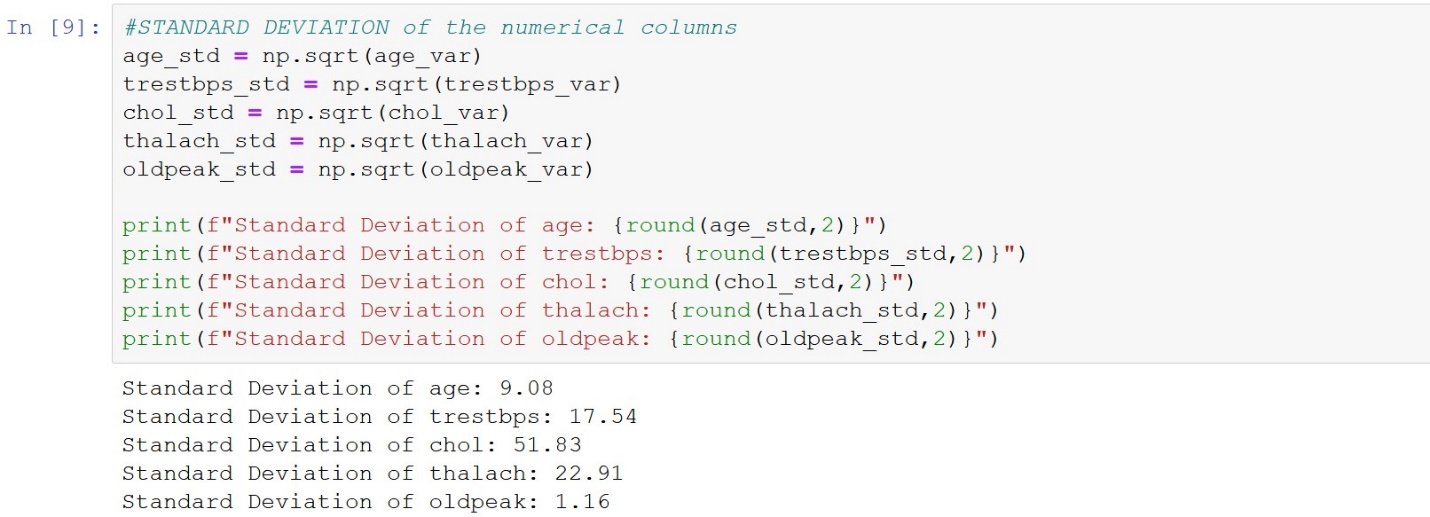
## Variance



To find the variance of the dataset we use the .var() function.

The variance is squared so it will not give us very precise information on the variability of data in the series. But, we will use it to find the distribution futher.

## Standard Deviation



To find the standard deviation of the data we need to square root our variance.

np.sqrt(variance) will give us the square root of variance which will be our standard daviation.

The standard deviation tells us how the the distribution of data around the mean value.

In our case the deviation from mean value are: -

Age = +/- 9.08

Trestbps = +/- 17.84

Chol = +/- 51.83

Thalach = +/- 22.91

Oldpeak = +/- 1.16

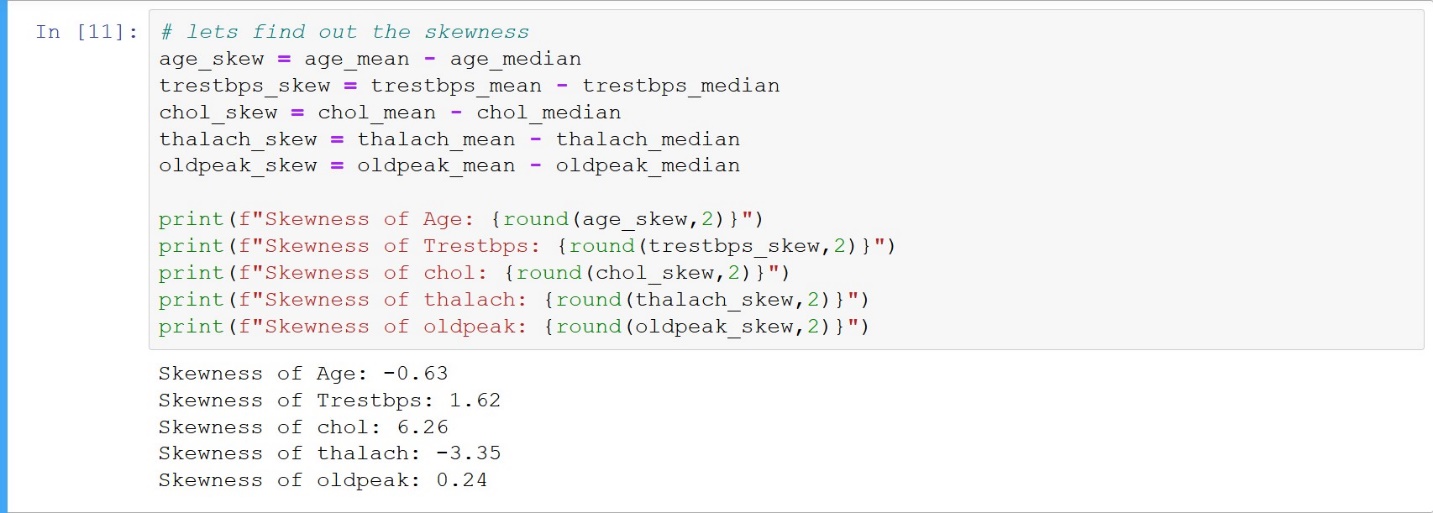
## Median



Median is the middle term in every data series. Which will help us to find the skewness of our distribution.

The np.median() function will calculate the median term from the every data series respectively.

## Skewness



Skewness of the data series will tell; which side are the outliers situated in the distribution. If the mean > median we say it is a positive skew and the outliers are situated at the right side of the distribution. And if the mean < median then it is a negative skew and the outliers are at the left side of the distribution.

Here we subtract the median value from the mean value to find the skew

Let’s check the skewness

Age – -0.63; it is negative skew but not high so there are very less outliers and situated at the left of the distribution.

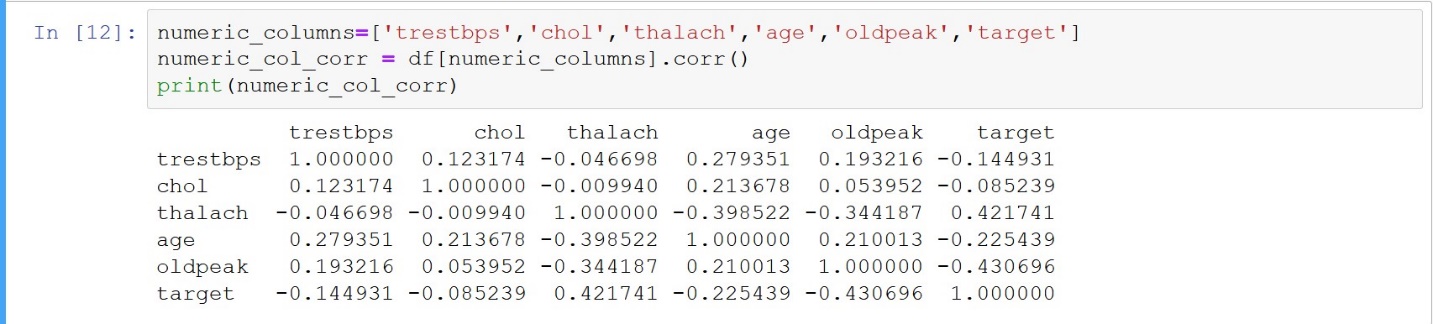
Trestbps – 1.62; it is positive skew but not high so there are very less outliers and at the right of the distribution.

Chol – 6.26; it is a positive skew and situated at the right of the distribution.

Thal – -3.35; it is a negative skew and situated at the left of the distribution.

Oldpeak – 0.24; it has the highest skew of all and it is positive so the outliers are situated at the right of the distribution.

# Correlation of numerical columns with target column



Created an array of all numeric columns and the target column as ‘numeric column’. The we find he correlation between all those columns with the target column and display correlations.

To analyze the data, we use heatmap

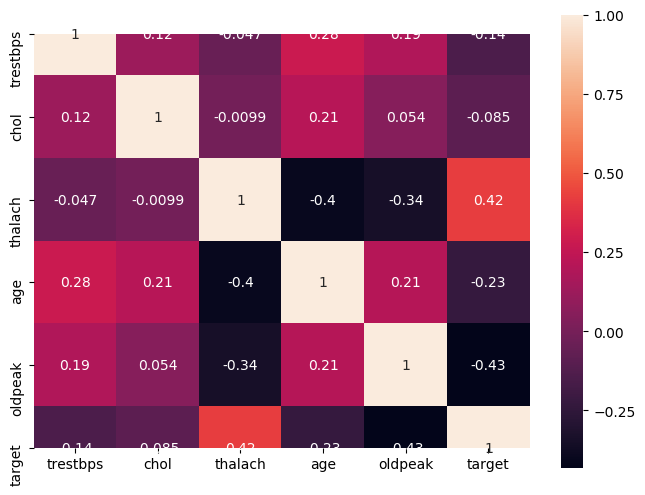


The figure function will create a canvas or 8,6 and dpi will give the resolution 100.

We create the heatmap of numeric\_col correlation i.e numeric\_col\_corr. The correlation will be labeled on the heatmap.

## Heatmap

Let’s see the heatmap and find good correlation.



In this heatmap we see that thal or maximum heart rate achieved has the highest positive correlation with our target i.e. 0.42. and old peak has the highest negative correlation i.e. -0.43.

This tell us that people with high maximum heart rate achieved will have significant impact on our having a heart disease or not. And ST depression induced by exercise relative to rest or oldpeak has very less impact on getting a heart disease or not.

# EDA of Numerical Columns

## Code block

sns.set\_style(“whitegrid”) – sets the graph background grids

plt.figure(figsize=(10,4),dpi=150) – creating a canvas height = 10 and width = 4, and resolution of 150

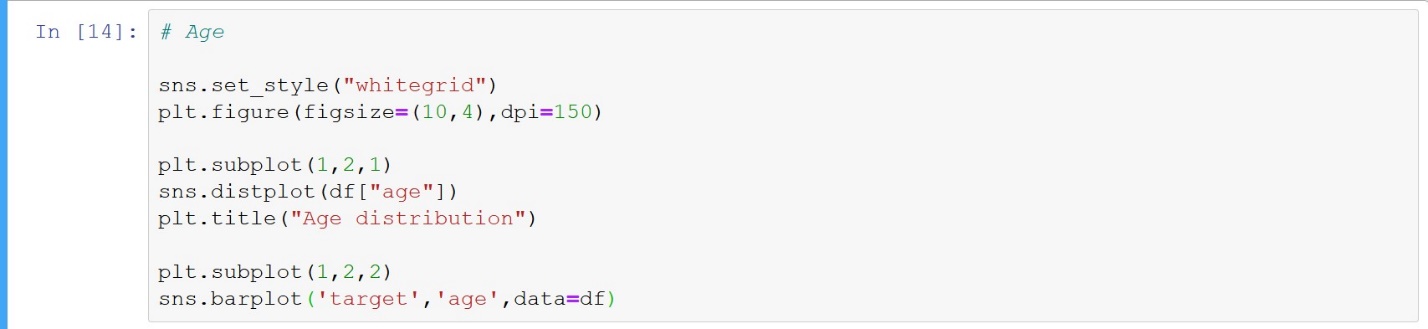
plt.subplot(rows=n,col=n,plot\_number) – subplot is used for creating different plot on the same canvas

sns.distplot(DF[“col\_name”] – creating histogram of the numerical column.

sns.barplot(x=dependent\_value,y=independent\_value,data=DF) – barplot displays the relationship between a dependent and independent column.

plt.title(“ “) – gives title to the particular subplot

## Age column





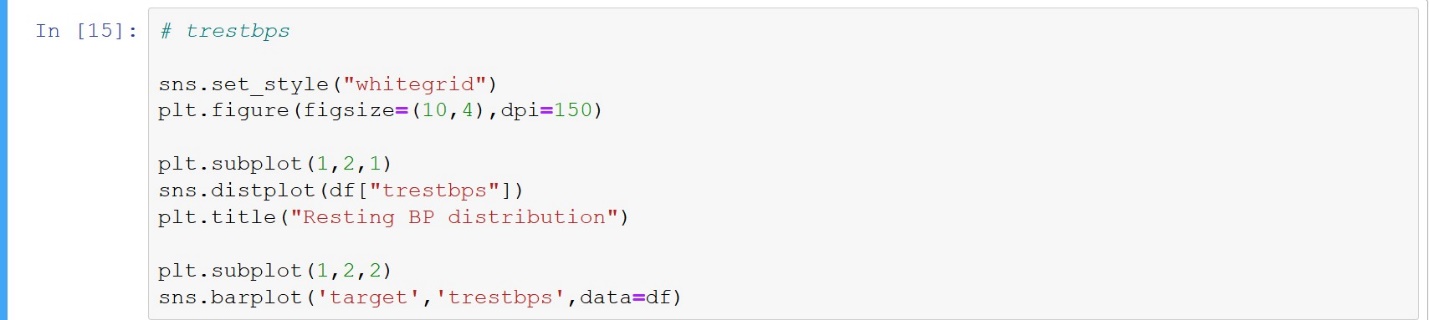
Subplot 1 – distribution plot.

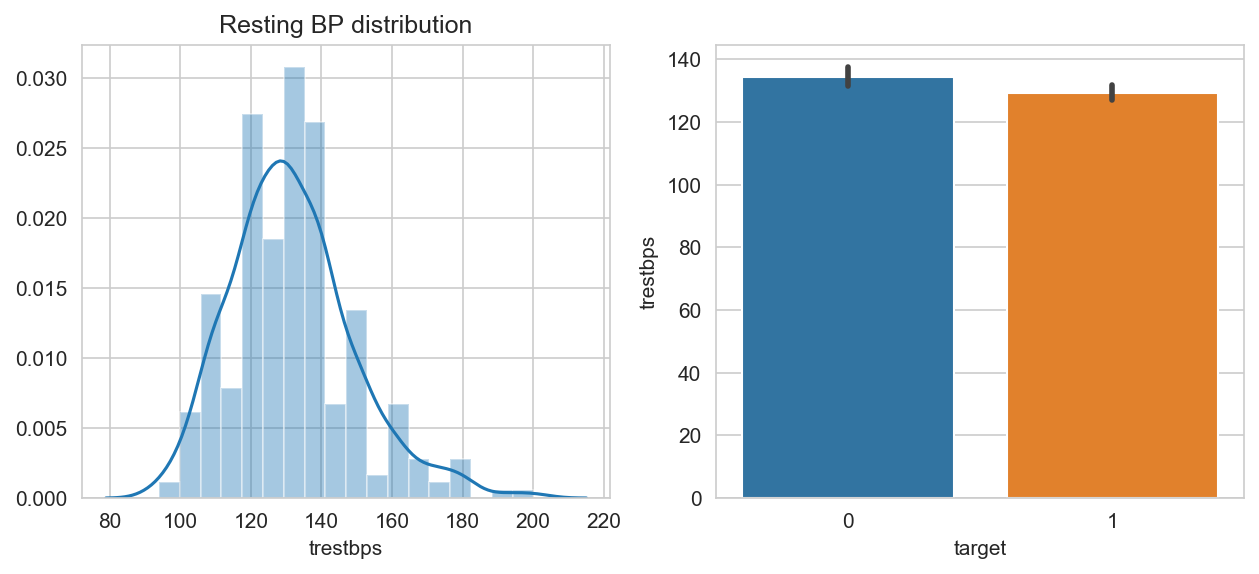
Most age group lies between 50 – 60. The skewness of the distribution is negative as we have seen skewness calculation, we have more of the higher age persons in our data set.

Subplot 2 – barplot

Most of the people from age 50 – 60 do not have heart disease. Whereas people from age 45 – 55 have heart disease.

## Resting Blood Pressure





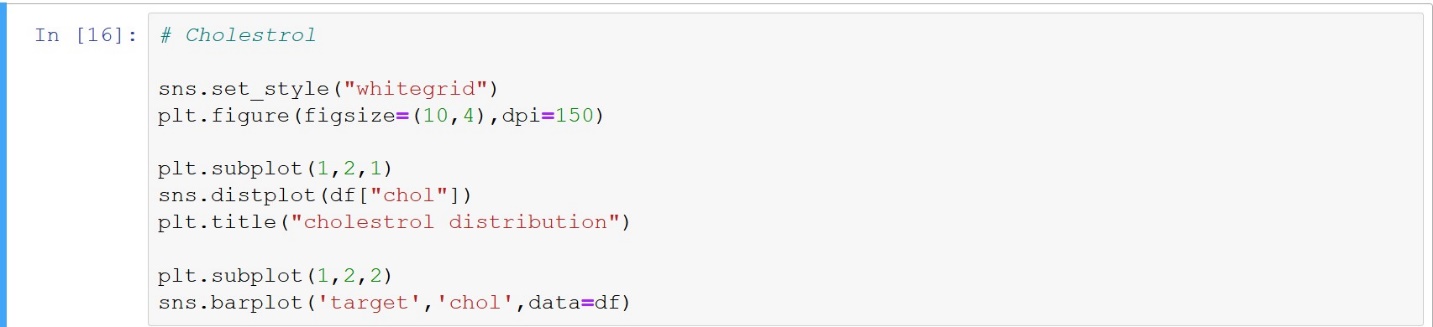
Subplot1 – distribution plot

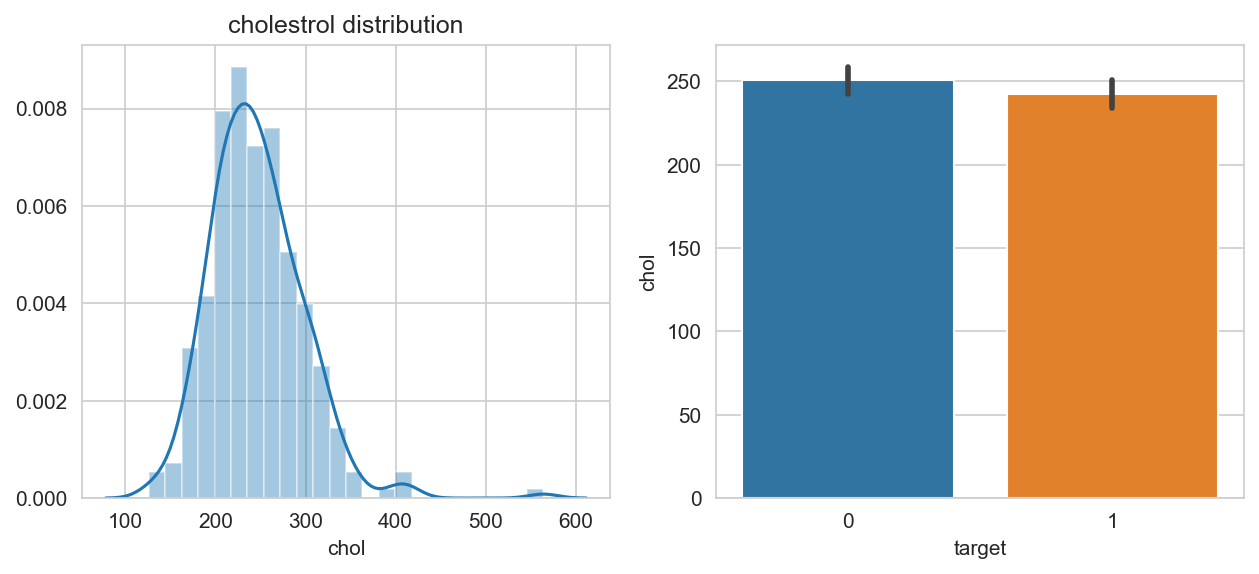
Resting blood pressure lies between 122 – 140. And it is right skew plot.

Subplot 2 – barplot

People with low resting blood pressure are more prone to have a heart disease.

## Cholesterol





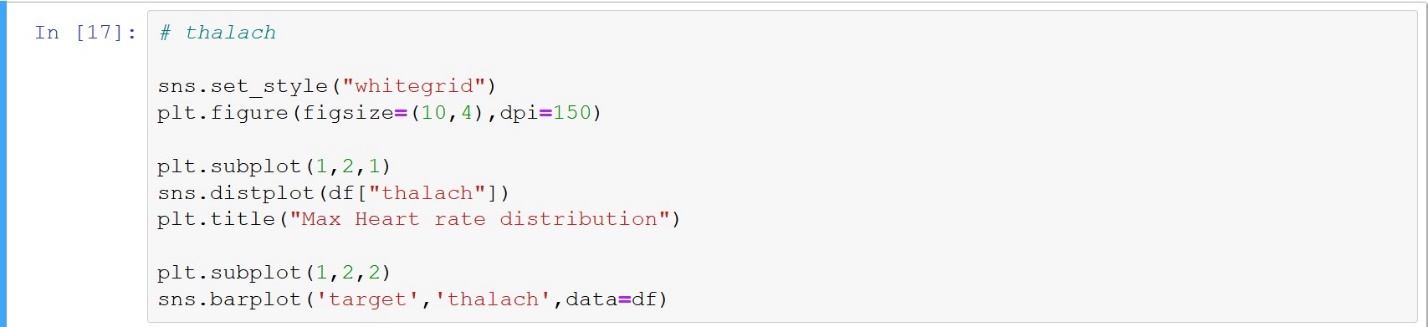
Subplot 1 – distribution

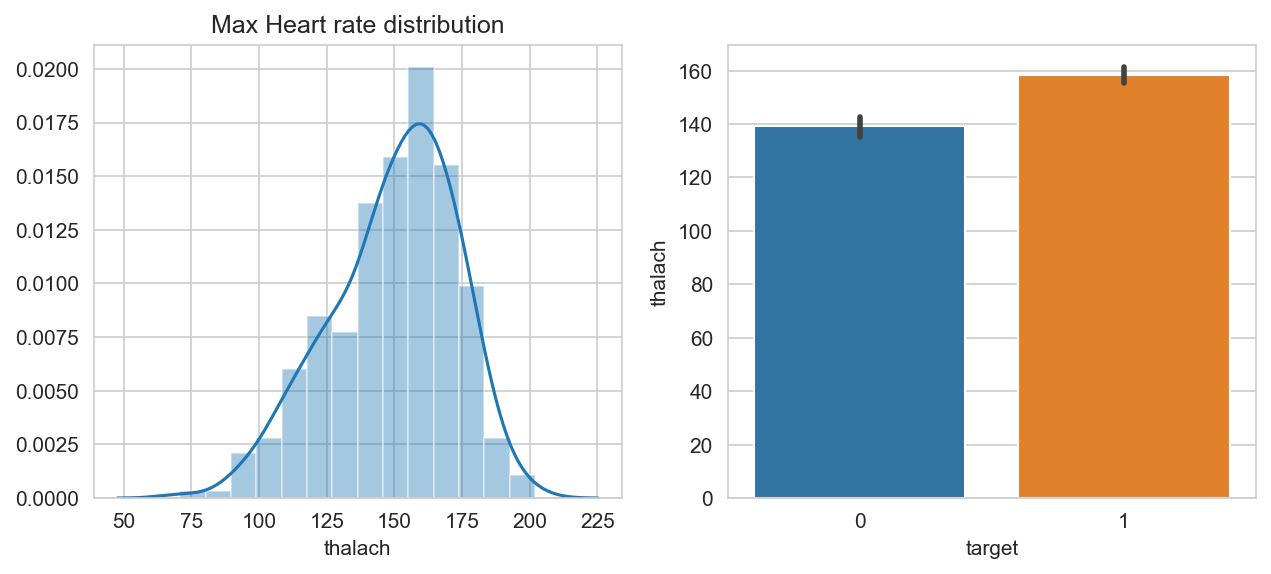
Average cholesterol level of person lies between 200 mg/dl – 250mg/dl. There are very few people with cholesterol level ranging between 540mg/dl – 600mg/dl, these are the outliers.

Subplot 2 – barplot

People with less cholesterol mg/dl have are more prone to have a heart disease.

## Maximum heart rate achieved





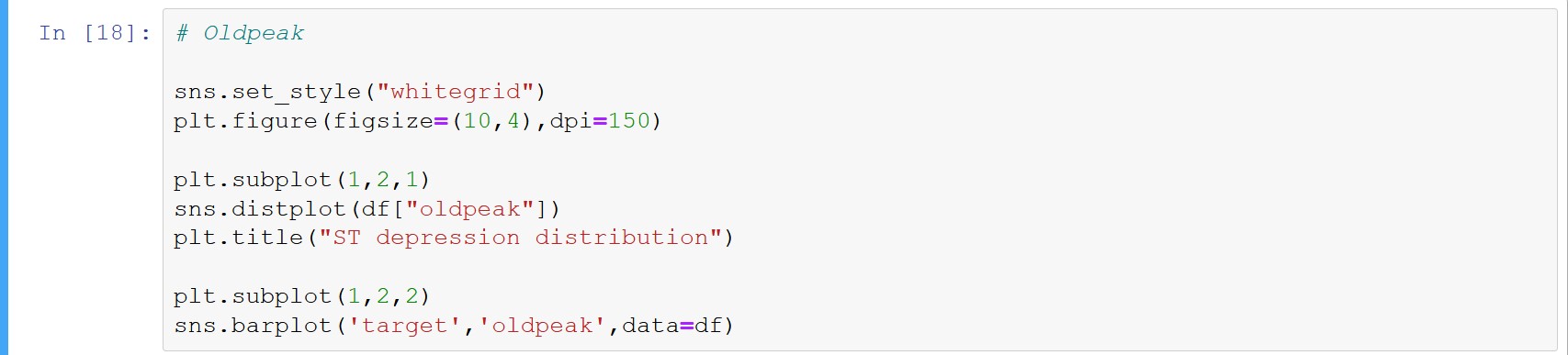
Subplot 1 – distribution

Average maximum heart rate achieved lies between 135 – 175. The distribution is left skewed which means few people are able to achieve less maximum heart rate.

Subplot 2 – barplot

People with high maximum heart rate achieved are more prone to have a heart disease.

## ST depression rate



Subplot 1 – distribution

Average ST depression lies between 0.1 – 1.4. It is right skew plot with high amount of outliers.

Subplot 2 – barplot

People suffering with heart disease are having low ST depression rate.

# EDA of Categorical Columns

## Code Block



sns.set\_style(“whitegrid”) – sets the graph background grids

plt.figure(figsize=(9,6),dpi=100) – creating a canvas height = 9 and width = 6, and resolution of 100

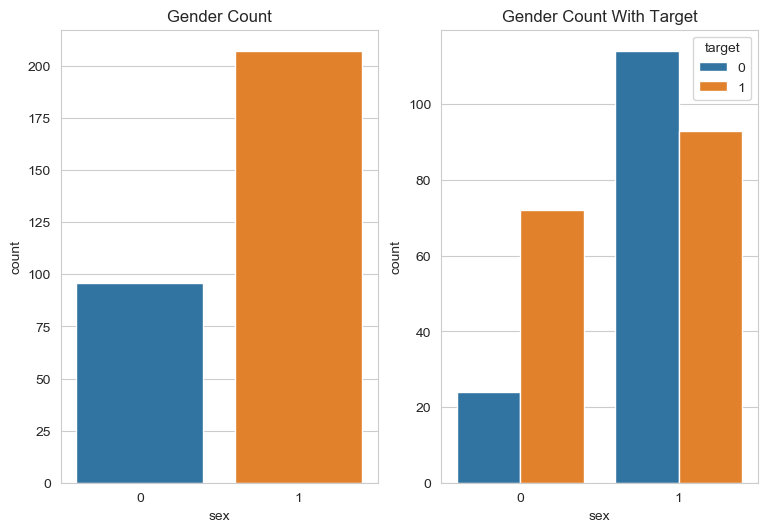
plt.subplot(rows=n,col=n,plot\_number) – subplot is used for creating different plot on the same canvas

sns.countplot(x=dependent\_value, data=DF,hue = Traget\_col) – countplot displays the count of each value in a categorical coumn.

plt.title(“ “) – gives title to the particular subplot

## Gender



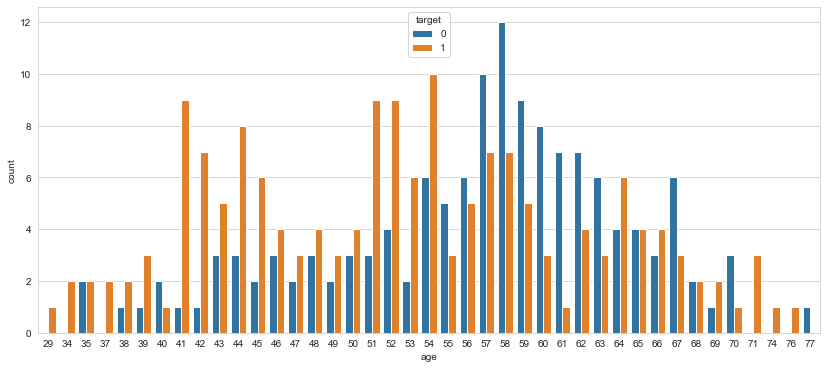


Subplot 1 – male = 1; female = 0

We have more male than female in our dataset

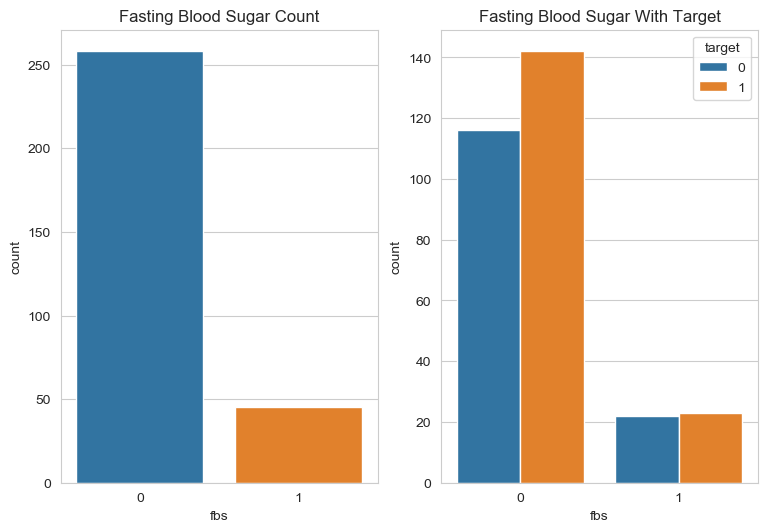
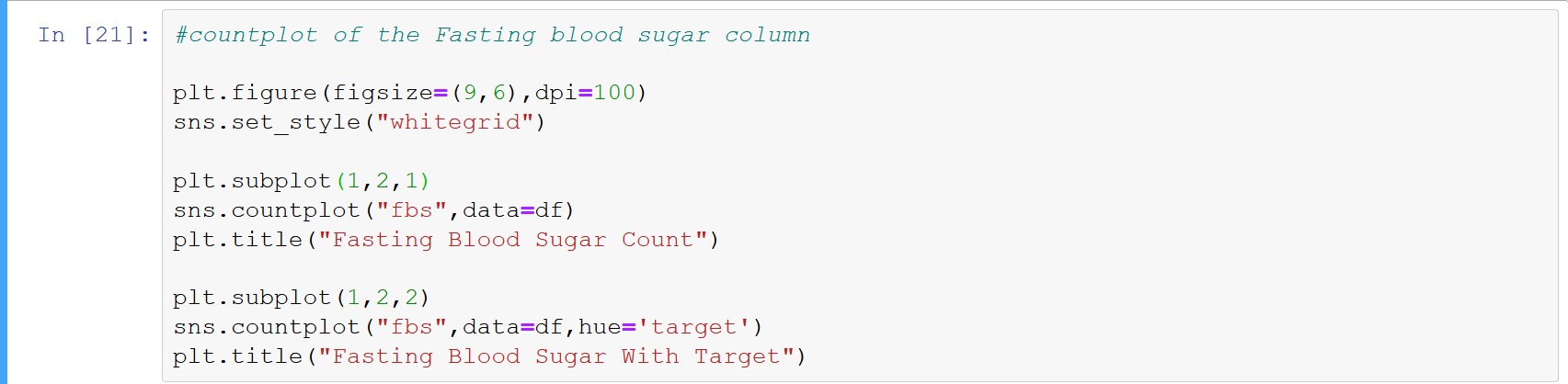
Subplot 2 – As per our dataset number of females suffering with heart disease more than males.

## Age with target



Form this countplot we see that age group of 41,42,44,45,51,52,53 and 54 most of the people are suffering from heart disease, whereas age groups of 57,58,59,60 and 61 are not suffering from heart disease.

## Fasting blood sugar



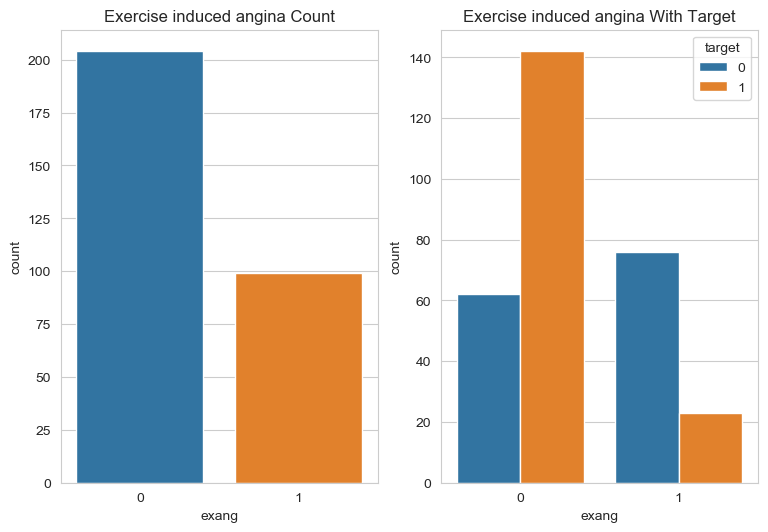
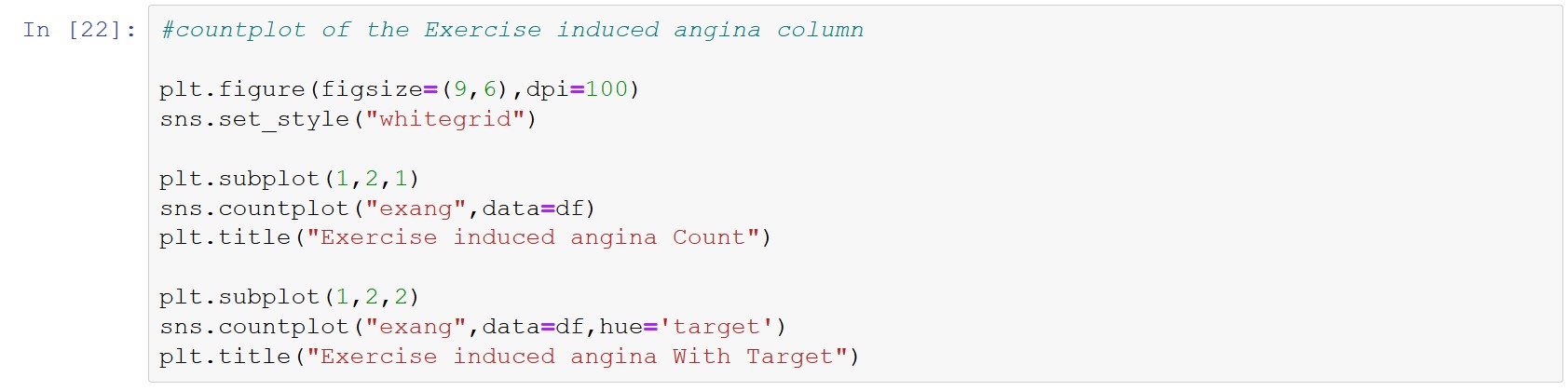
Subplot1 – people with fasting blood sugar = 1; people without fasting blood sugar = 0

Here we see that most of the people do not have fasting blood sugar.

Subplot 2 – people having fasting blood sugar are equal in number of having and not having a heart disease.

Whereas, people who do not have fasting blood sugar and suffering from heart disease is more than that of people not having fasting blood pressure and not suffering from heart disease.

## Exercise induced angina

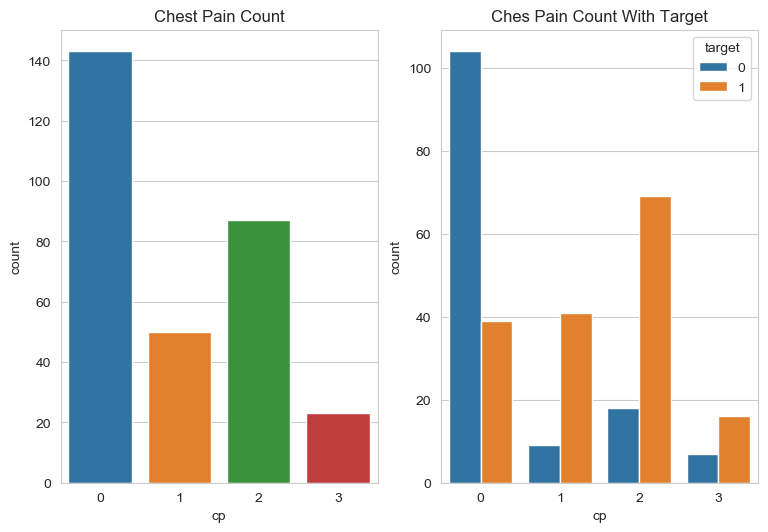
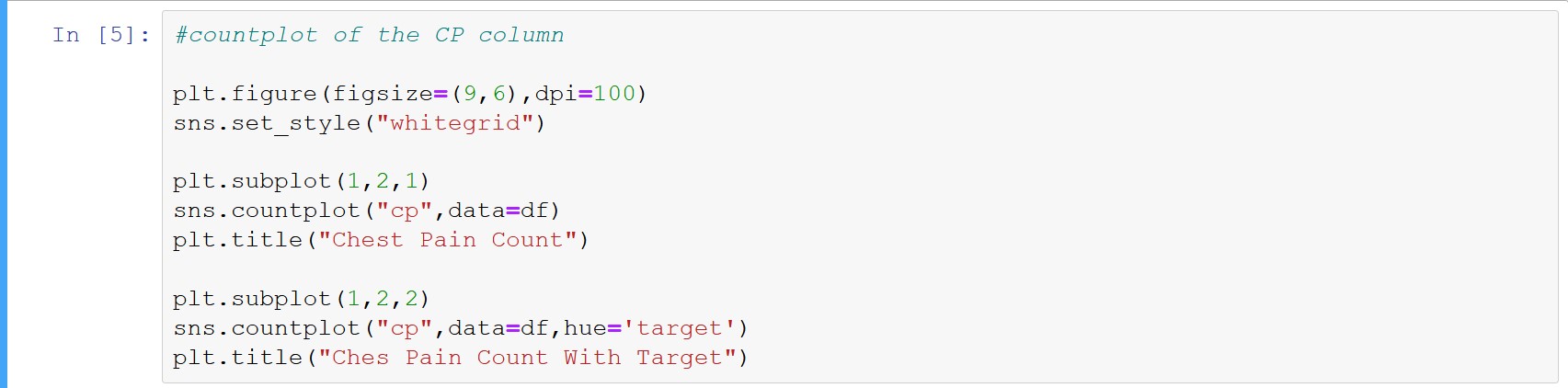


Subplot 1 – having angina = 1; not having angina = 0

People having exercise induced angina is less than that of people not having exercise induced angina.

Subplot 2 – most people having the angina do not have heart disease. Whereas, people not having the angina and suffering from a heart disease is more.

## Chest Pain

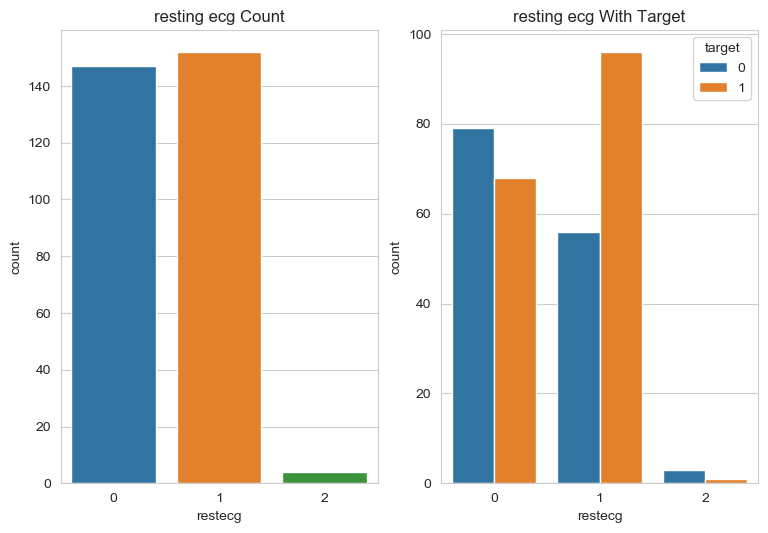


Subplot 1 – 0 = typical angina; 1=atypical angina;2=non-anginal; 3=asymptomatic

Most of the people and typical angina and very less from asymptomatic

Subplot 2 – most of the people with typical angina do not have heart disease. People with atypical and non-anginal chest pain are more who are suffering from heart disease.

## Resting electrocardiographic result

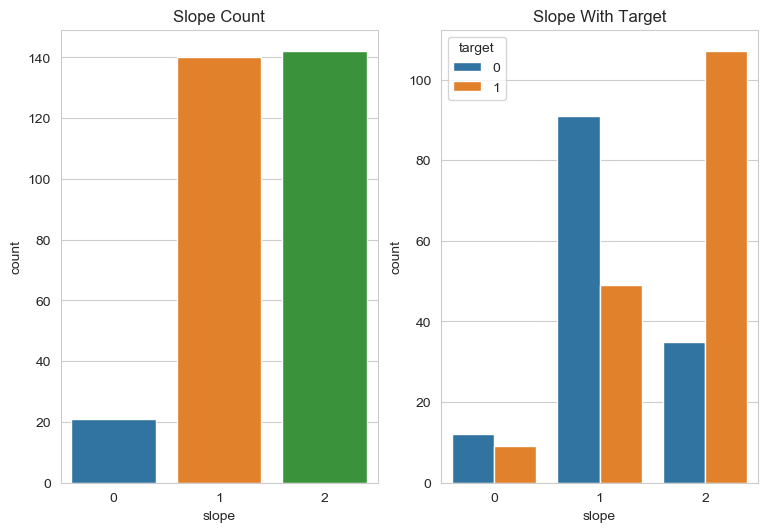


Subplot 1 – resting ecg has three values 0, 1 and 2

Most of the people lie on value 0 and 1 of ecg and very less for 2

Subplot 2 – more people with resting ecg of value 1 are suffering from heart disease. less people with value 0 have a heart disease. If we look at the vale 3 ecg only 1 or 2 people suffer from heart disease and most don not.

## ST Segment

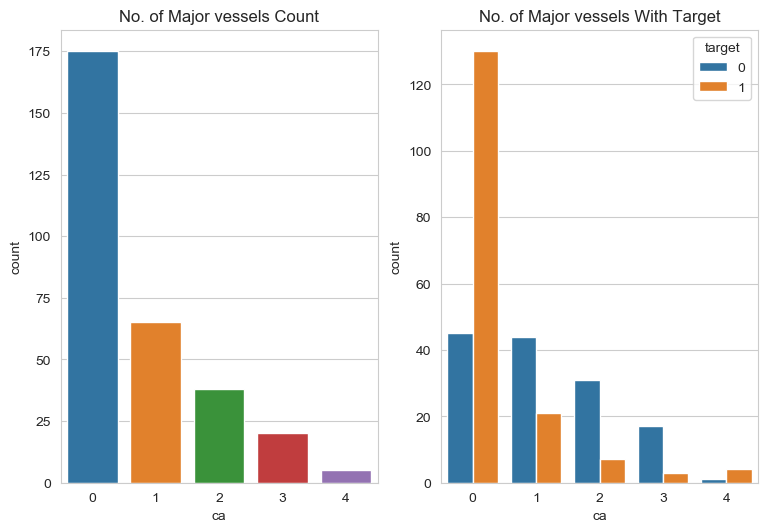


Subplot 1 – 0 = upslopping; 1= flat; 2=downslopping

Most of the people have an upslopping or downsloping ST segment

Subplot 2 – there are more people with a downslopping ST segment that suffer from heart disease. Whereas less people with flat ST segement have heart disease.

## Major blood vessels

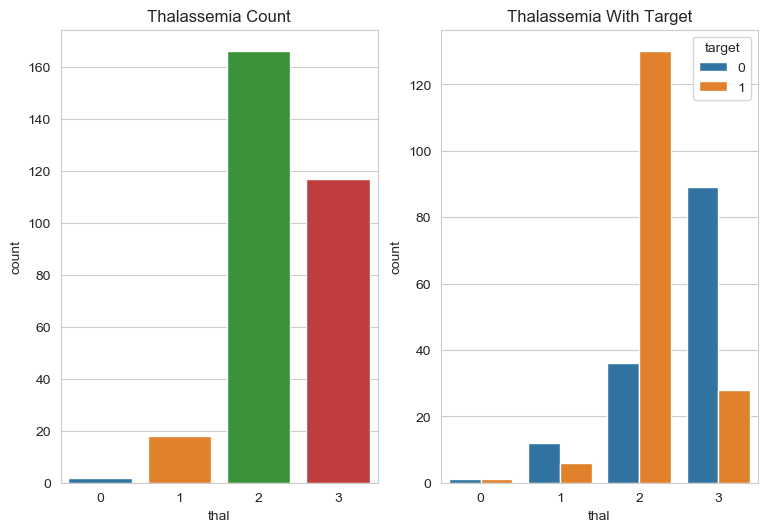


Subplot 1 – major blood vessels numbered as 0,1,2,3 and 4

Most people having 0 and 1 blood vessel whereas 4 is very less.

Subplot 2 – more people are suffering from heart disease with 0 and 4 if we see the count of 4. And for 1, 2 and 3 having heart disease is less.

## Thalassemia



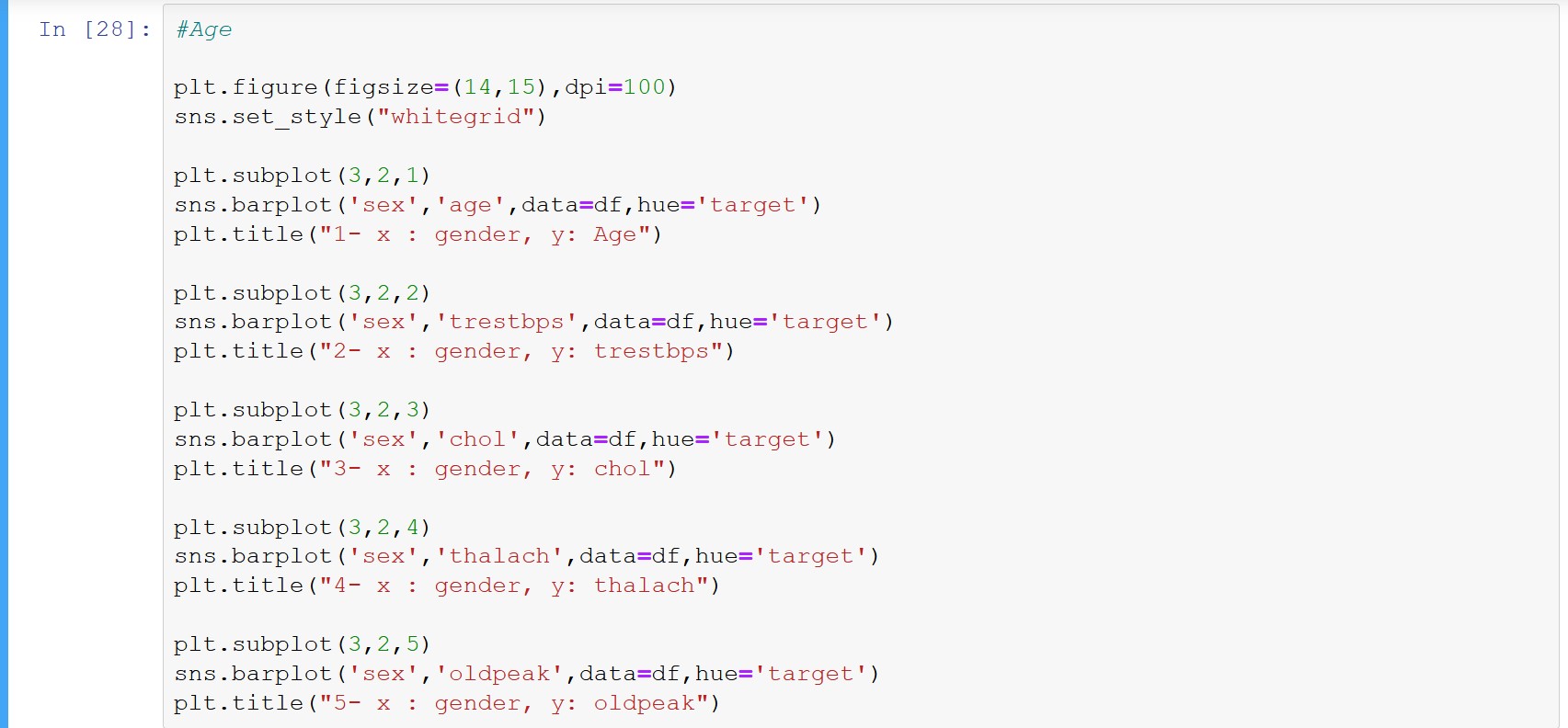
Subplot 1 – thalassemia and its condition valued from 0 – 3 in increasing order.

Most people suffer from thalassemia of 2 order and 3 order

Subplot 2 – more people with 2 order are having heart disease and less people of 3 order are having the heart disease.

# Relation b/w Numerical & Categorical Column w.r.t Target

## Code block



sns.set\_style(“whitegrid”) – sets the graph background grids

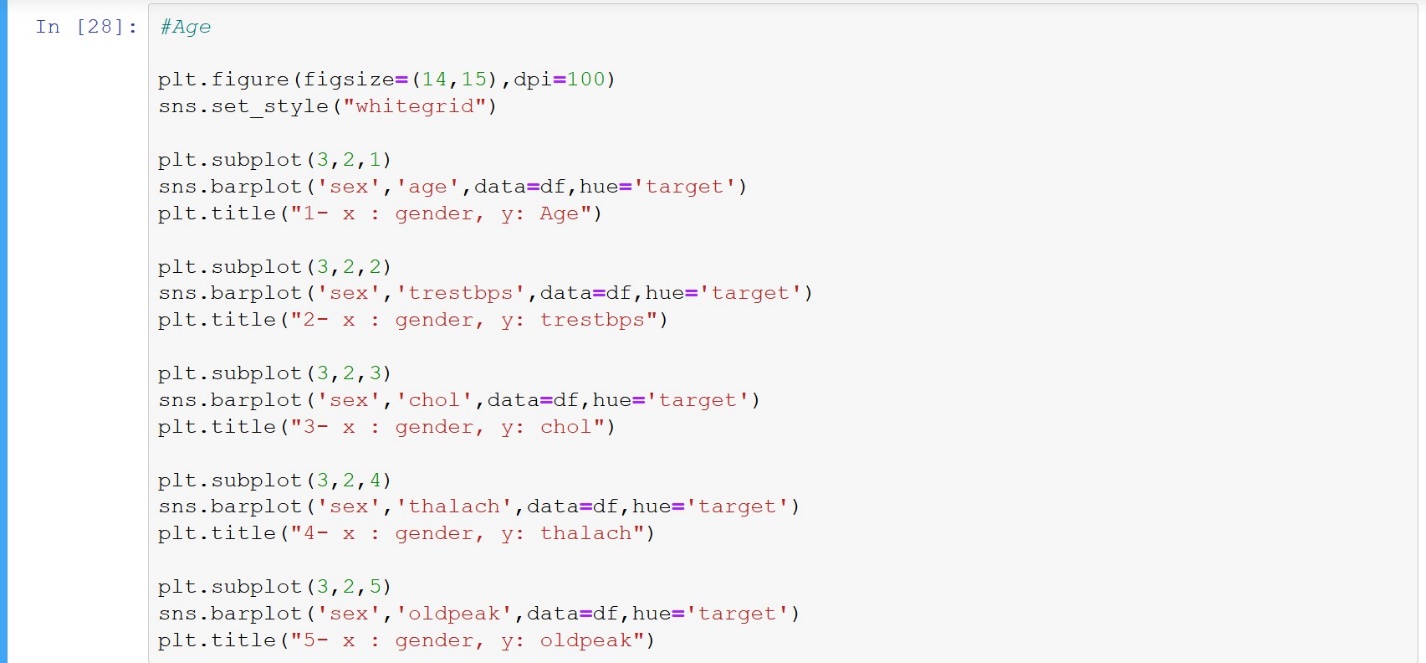
plt.figure(figsize=(14,15),dpi=100) – creating a canvas height = 14 and width = 15, and resolution of 150

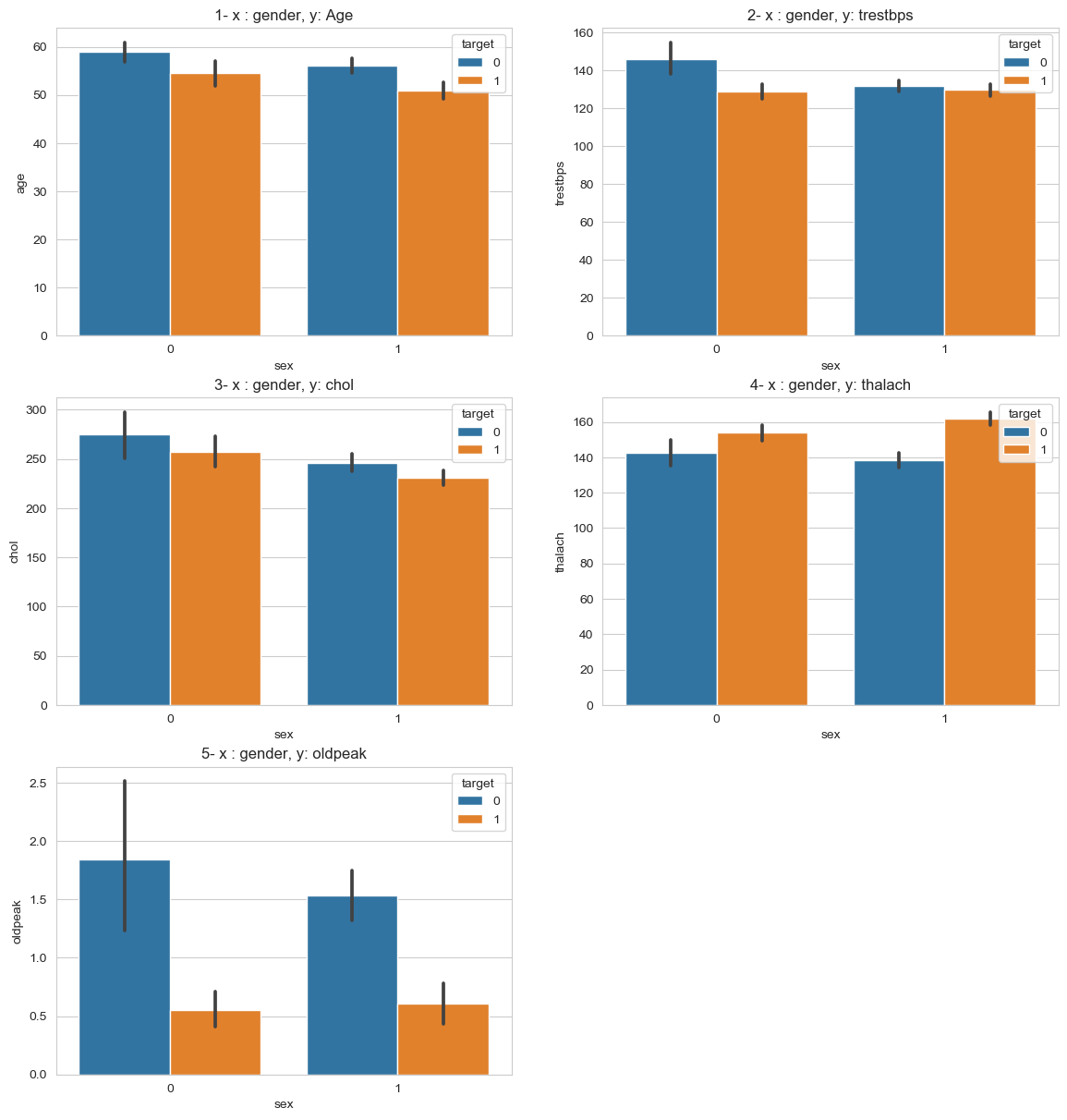
plt.subplot(rows=n,col=n,plot\_number) – subplot is used for creating different plot on the same canvas

sns.barplot(x=dependent\_value,y=independent\_value,data=DF,hue=Target col) – barplot displays the relationship between a dependent and independent column. Hue argument divides the barplot according to the target column.

plt.title(“ “) – gives title to the particular subplot

## Age and Numerical columns





1 – in this dataset, number of elderly females are more than of elderly males. Age wise, people suffering from heart disease is almost same in both the gender. As per block 19, count of males are more. So here we can say males suffer from heart disease more than females.

2 – for males the resting blood pressure does not matter that much for having or not having the heart disease but for females with high resting blood pressure there are less chances of having a heart disease. But resting blood pressure for having a heart disease is almost equal for both males and females.

3 – cholesterol mg/dl is higher for females. As per block 16 people suffering from heart disease have low cholesterol value. Here we see males have low cholesterol to that of females.

4 – Maximum heart rate achieved is major cause of heart disease for both male and females.

5 – the slope of the peak exercise ST segment in females with no heart disease is very high compared to that to males.

## Fasting blood sugar with numerical columns



1 – People with fasting blood sugar have higher chances to having a heart disease. People with and fasting blood sugar and heart disease are slightly elder than people not having fasting blood sugar.

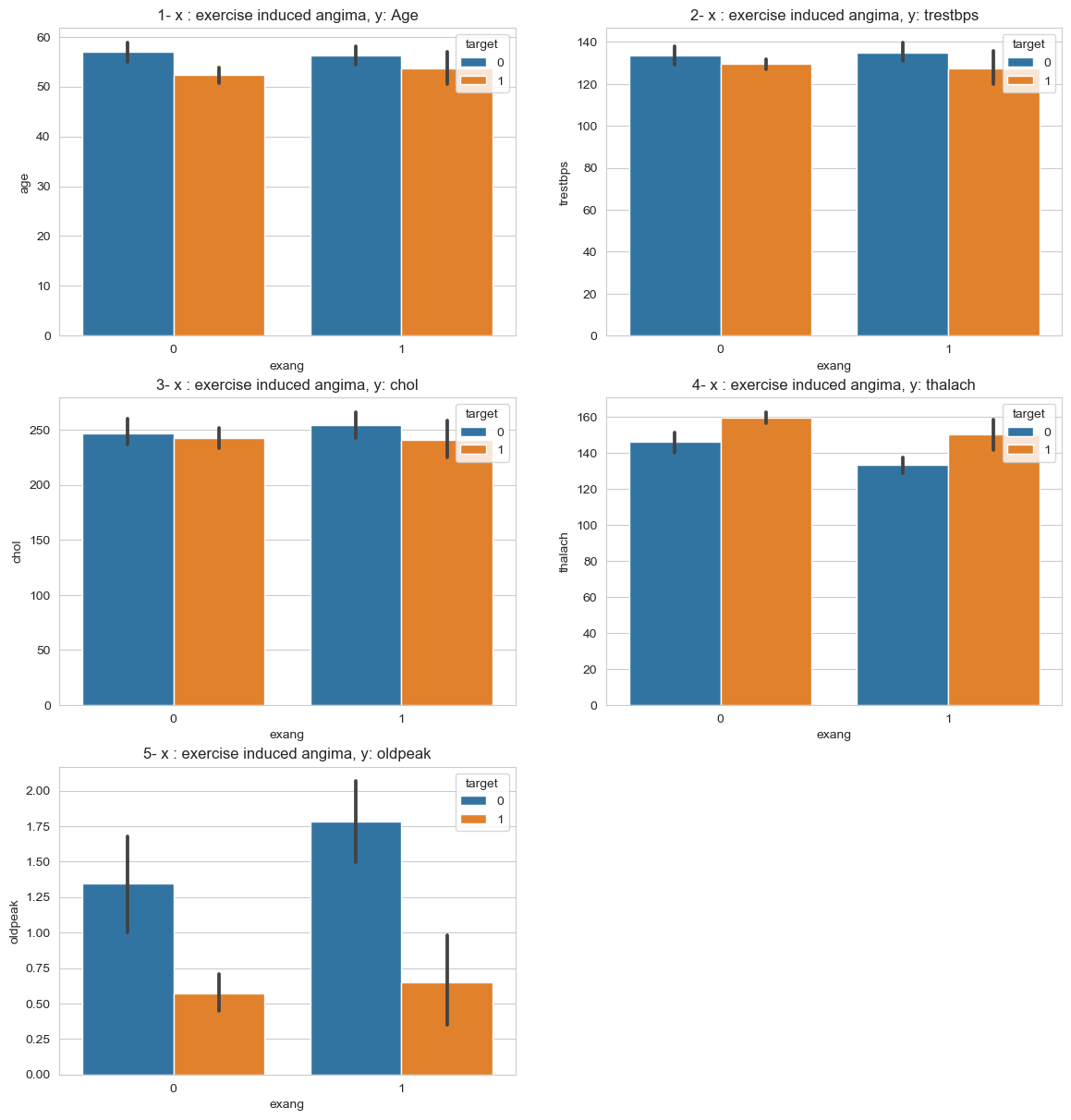
2 – people without fasting blood sugar and no heart disease have lower resting blood pressure than that of people with fasting blood sugar.

3 – people with fasting blood sugar but not heart disease have high cholesterol mg/dl. as we have seen earlier people with heart disease tend to have low cholesterol, here we see that people with fasting blood sugar and low cholesterol have heart disease.

4 – high maximum heart rate achieved is a major cause for heart disease but in case of fasting blood sugar people with or without fasting blood sugar have almost same maximum heart rate achieved.

5 – there are many outliers situated where people have fasting blood sugar but not heart disease. People with fasting blood sugar and heart disease is low but again there are many outliers. So we say that there are many people with fasting blood sugar and heart disease but there ST depression distribution is high.

## Exercise induced angina with numerical columns



1 – people with or without exercise induced angina both lie the same age range of having or not having heart disease

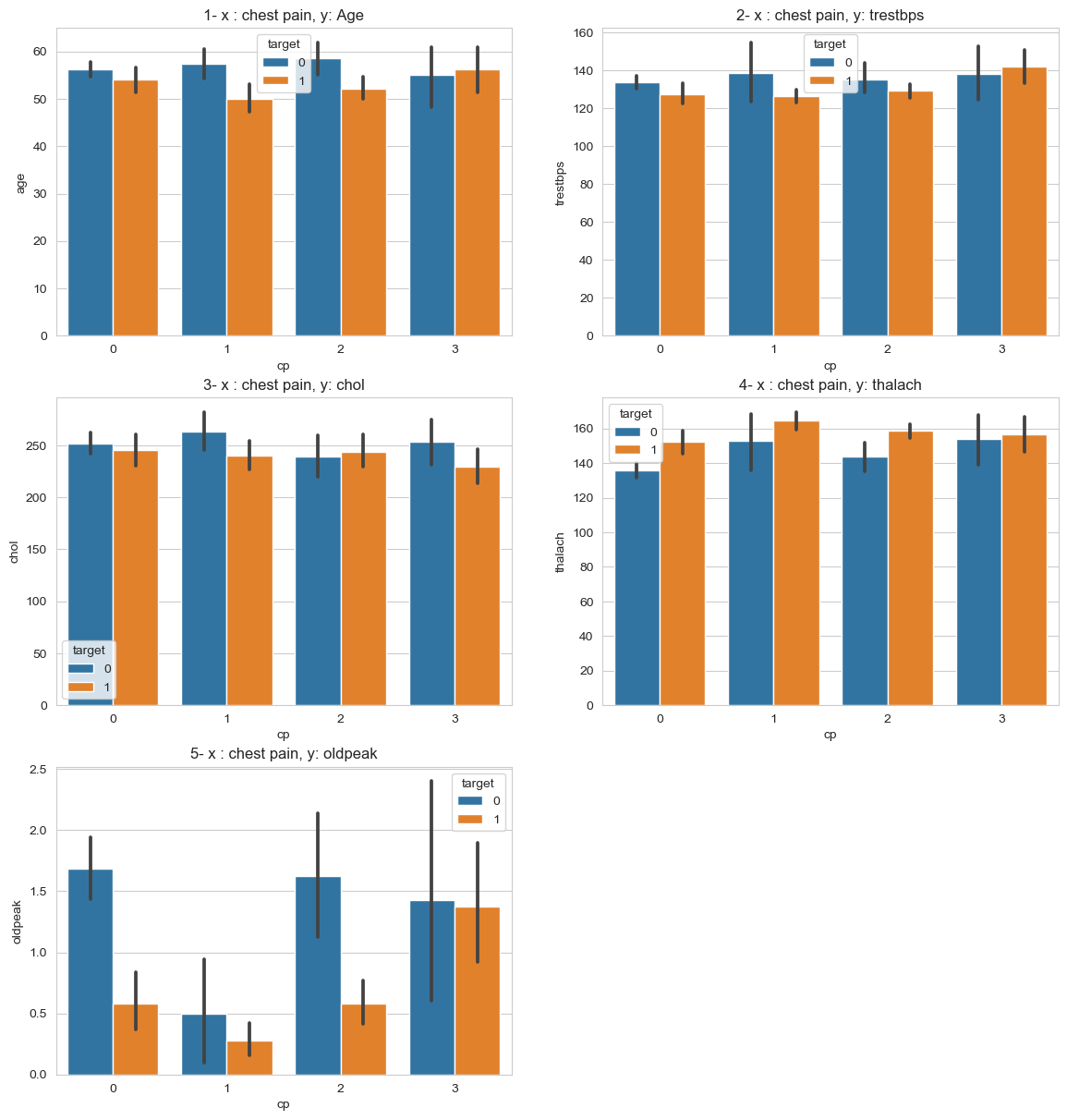
2 – people with or without exercise induced angina both their resting blood pressure lie in the same range having or not having heart disease.

3 - people with or without exercise induced angina both their cholesterol mg/dl values lie in the same range having or not having heart disease

4 – people with exercise induced angina and suffering from heart disease have low maximum heart rate achieved from that of people without the angina and suffering from heart disease.

5 – the ST segment for the people with the angina but not heart disease is very high. So we can say that people how have good ST segment will get high maximum heart rate but can have less chances of suffering from heart disease.

## Chest pain with numerical columns



1 – people with asymptomatic chest pain and having heart disease are older in age. People with atypical and non-aginal chest pain and suffering from heart disease are older.

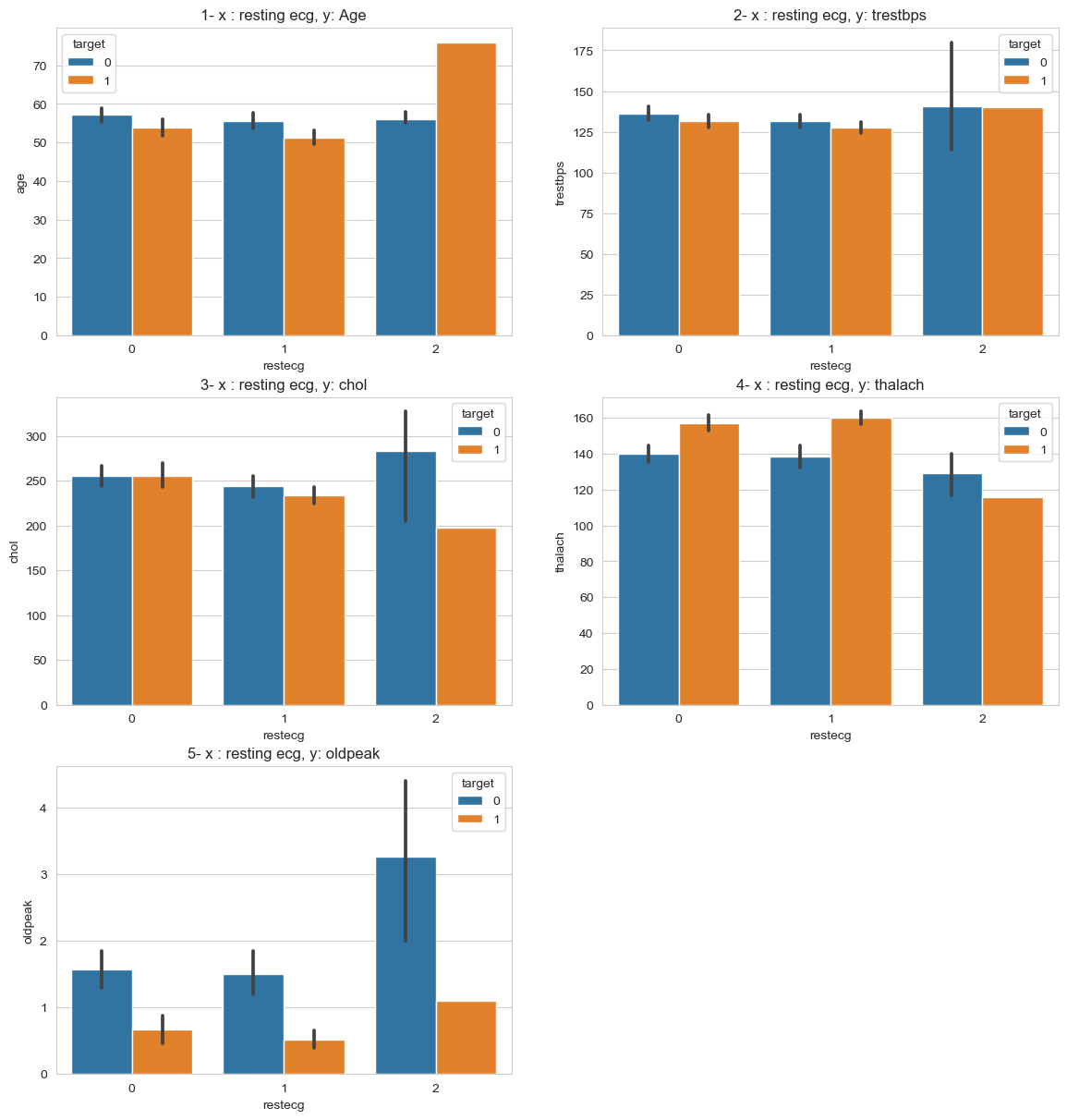
2 - people with asymptomatic chest pain and suffering from heart disease tend to have high resting blood pressure. And people with atypical and non-aginal chest pain but not suffering from heart disease is have higher resting blood pressure.

3 – cholesterol mg/dl for every type of chest pain and having heart disease is almost the same. We can say cholesterol in correlation with heart disease has a very less impact.

4 – people with atypical chest pain and high maximum heart rate achieved has more chances of getting heart disease. More people are suffering with typical heart disease but those who have high maximum heart rate tend to have heart disease.

5 – people with asymptomatic chest pain who are having and not having heart disease have high but almost same ST depression. And rest three chest pain type where people are not having heart disease their ST depression is high.

## Resting electrographic with numerical columns



1 – people with rest ecg of order 3 are and having heart disease is above the age of 70, and those with no heart disease lie between 50 – 60. Order 1 and 2 both having and not having heart disease lie between age 50 – 60.

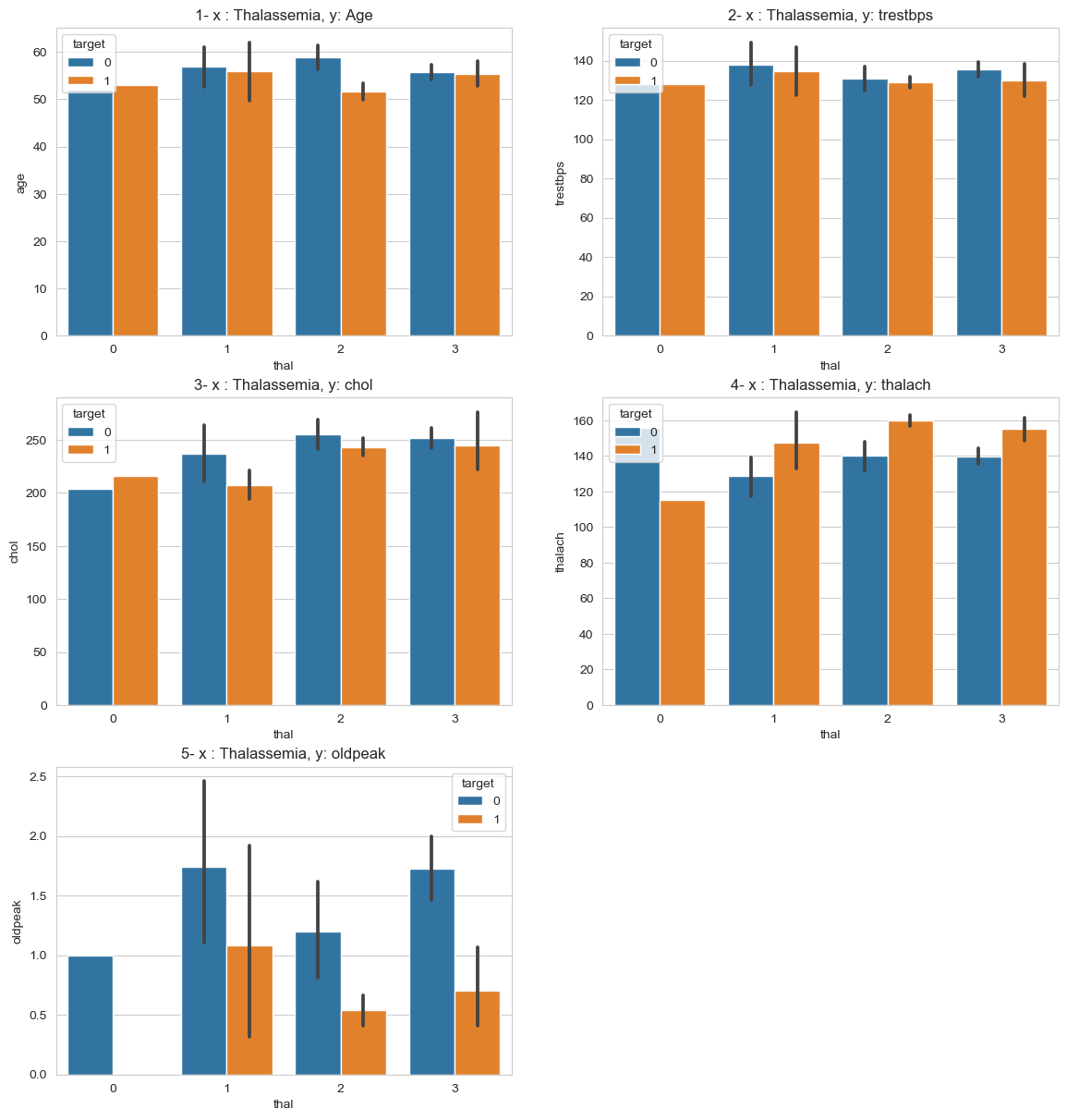
2 – people with rest ecg order 3 suffering and not suffering from heart disease have equal resting ecg, though not suffering people have some outlier. Order 1 also has the same resting ecg for both suffering and not suffering heart disease but lower.

3 – people with high cholesterol and order 3 rest ecg do not have heart disease, and those who are suffering in order three have very low cholesterol.

4 – people with order 1 and 2 rest ecg who are having high maximum heart rate achieved suffer from heart disease.

5 – most people with order 3 rest ecg having high ST depression do not suffer from heart disease. Order 3 people have highest chances out of all three to rest ecg to have heart disease.

## Thalassemia with numerical columns



1 – people with thalassemia of order 2 and not having heart disease are older than that or order 2 and not having heart disease. For order 0, 1 and 3 both having and not having heart disease lie in equal range of age.

2 – people with order 1 thalassemia and not suffering from heart disease have higher resting blood pressure, then that of suffering from heart disease. As most of the people are suffering from order 2 thalassemia we can say that people with high resting blood pressure do not suffer from heart disease.

3 – as most of the people are having thalassemia of order 2. We can say lower the cholesterol mg/dl level higher are the chances of getting a heart disease.

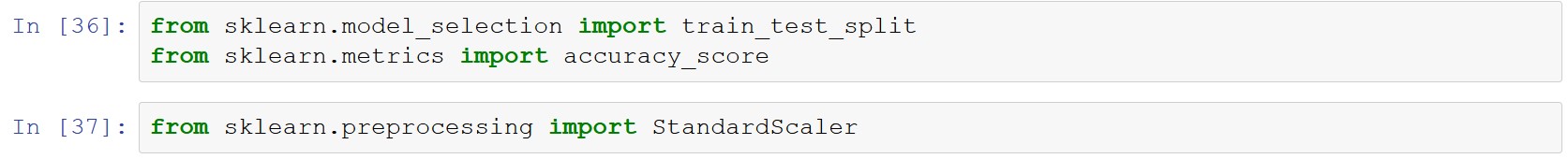
4 – maximum heart rate achieved for any order where people are not having heart disease is lower than that of people having heart disease.

5 – the ST segment for order 0 and having heart disease is almost zero. We can say people having the ST segment of 0 have very less chances to suffer from heart disease. And for order 1 there are comparatively high chances. For order 3 if the ST segment is high there are less chance for heart disease.

# Data Preprocessing



we drop the “target” column and take rest of the column an place them into variable “ x ”. “y” variable holds only the “target column”.



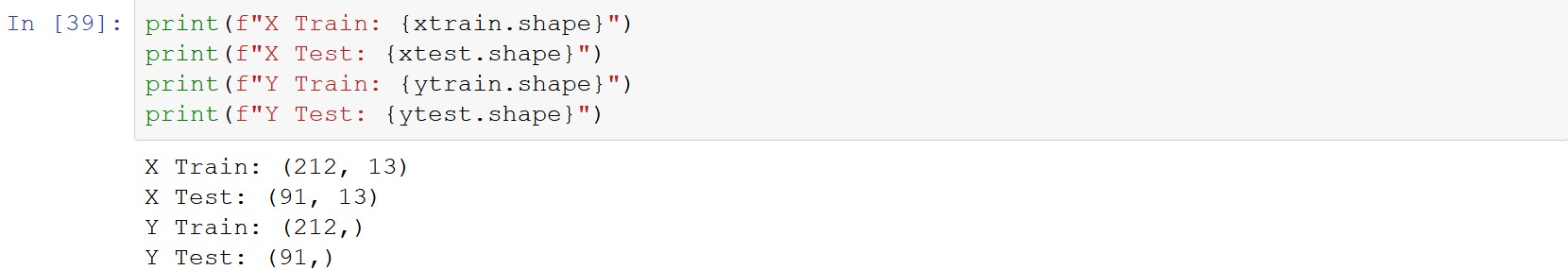
Importing train\_test\_split for splitting our data into training and testing model.

Accuracy\_score will tell us how accurately our model predicted our data.

Standard Scaler will scale for good results.



Tuple unpacking to create training and testing data set. Here we randomly select 70% data and place from both x and y and place them into xtrain and ytrain, remaining 30% will be placed in xtest and ytest to test the result on the trained data. Test\_size = 0.3 will do the dividing while unpacking. Random\_state = 101 just sets the random number which are divided between training and testing at their place not changing them every time we run the cell.



.shape will display the shape of the training and testing divided data sets.

# Machine learning prediction on the Dataset

## Decision Tree

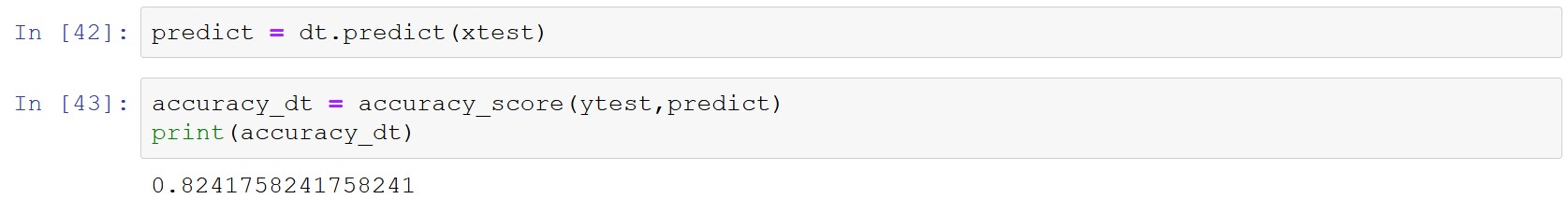


Importing Decision Tree Classifier from sklearn library



Instantiated Decision Tree Classifier in “dt”.

And fitted xtrain and ytrain into dt for training the model.



After training the model with xtrain and ytrain we predict the data from xtest and place all the predicted values in predict variable.

Then we check the accuracy of the predicted data by comparing the predicted data and actual data, and place it into “accuracy\_dt”.

And print the data.

Now we can pass our data to check the result.

## K-Nearest Neighbors



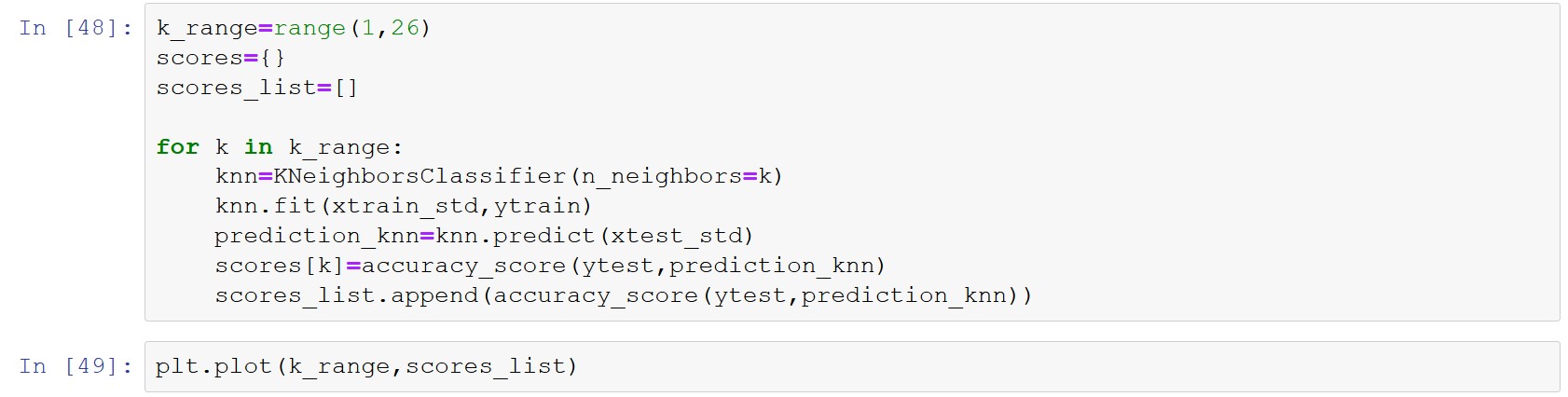
We instantiate StandardScaler into variable scaler.

Then we fit the xtrain and transforming it and placing it into xtrain\_std.

We just scale and transform the xtest and place it into xtest\_std.



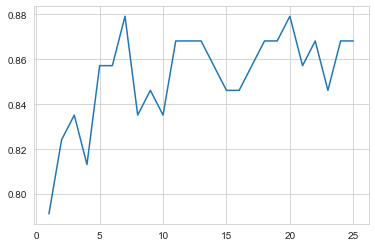
Importing K nearest neighbors



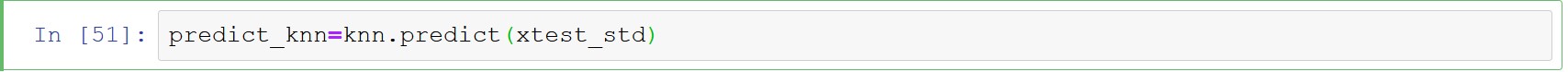
For good result with knn we need to give the number of neigbors we need to detect. For that we make a plot which shows at what number of neighbor will our model predict with good accuracy.

Then we plot the graph.

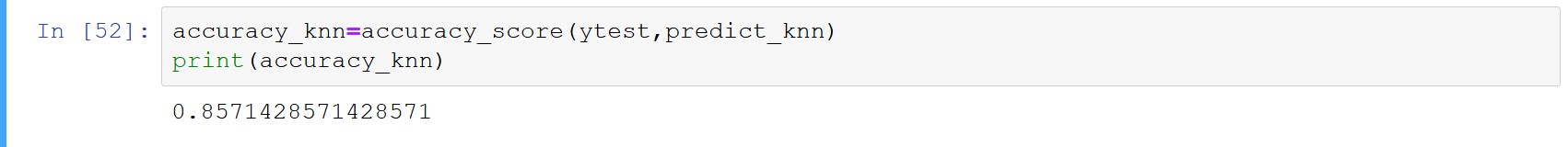
Here we see our model is predicting good at around 6 and then at 20. So we take least neigbors for more flexible model.



We instantiate the K-nearest neighbor classifier with number of neighbors = 6 to get good result.



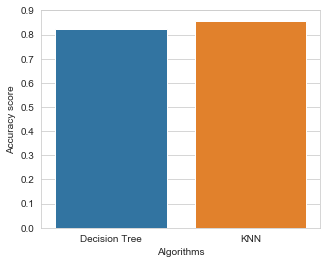
Predicting the values with knn.predict on xtest\_std



We check the accuracy of the knn prediction

Now we can say if we give our own set of values to our KNN model there is a 85% chances to predict it right.

# Comparing Accuracy between our KNN and Decision Tree Classifier



We see here that KNN is predicting our data with more accuracy.

# Conclusion

Heart diseases are one of the major concerns of society and the number of people affected by these diseases is increasing day by day and it is important to find a solution to this problem.

It is difficult to manually determine the odds of getting heart disease based on risk factors. But with the help of data analytics and machine learning models, we can determine these diseases and have a better chance of treating it.